# Corpus-Assisted Discourse Studies (CADS)

## Video 1 transcript

Full resource, see: <https://www.ncrm.ac.uk/resources/online/all/?id=20855>

Hello. My name is Maria Leedham, and I’m a Senior Lecturer in Applied Linguistics at the Open University. In this first video I’ll introduce the methodology of CADS and explore why you might use this.

CADS or Corpus-Assisted Discourse Studies, is a mix-method approach that combines the computational power of corpus linguistics with the interpretive insights of discourse analysis. So, it’s combining quantitative and qualitative methods to give a more powerful and rounded approach to your research. CADS starts from a social issue, for example CAD studies have been used to explore the language surrounding particular groups in society, such as refugees and asylum seekers, and what this might reveal about societal attitudes towards these groups. And work has been carried out in a variety of fields, for example, health communication, academic writing, TV shows, on line grooming, to name a few. In later videos you’ll look at some examples of CADS research taken from my own work on student writing, how social workers are portrayed in the press and on young adult fiction.

To start with, let’s look at what is meant by corpus linguistics. Corpus, from the Latin for body, refers to a systematically collected body of language generally stored computationally. So, what corpus analysts do is to compile a corpus or look at a pre-existing one and explore it using specialist software. And the advantage of using corpus linguistics is that you’re looking at a lot of text and, as John Sinclair says, the language looks rather different when you look ta a lot of it at once. So, collecting a large quantity of text together means you can slice through the data in different ways and get fresh perspectives. And you’ve still got sight of the individual texts so you can zoom in and out and look at them at different levels. And seeing a lot of language within a lot of text at once gives you this fresh perspective to see things differently and maybe notice things you might not see if you were to read one text at a time.

Corpora now range in size from billions of words scraped from webpages, to small, carefully formed corpora, for example, 10,000 words of Facebook updates from particular individuals. The size of the corpus should relate to your research aims. Maybe a large general reference corpus would do for what you want to find out. But if you’re looking for something more bespoke, for example how social workers are portrayed in UK newspapers, then you’ve probably got to build your own corpus.

On this slide I discuss ways of analysing a corpus, so these are all entry points into your dataset. The first bullet point lists descriptive statistics, for example looking at the mean average sentence length, which might be useful if you’re looking at learner language. You would expect more advanced students to have longer mean average sentence lengths in their written work when compared to beginner students, and the same with a variety of lexical items. Is their vocabulary perhaps more varied, a wider range of items?

Frequency lists are ordered lists from most to least frequent words in a corpus. And you might compare one of these with a larger reference corpus, and we turn term these keyword lists, and that’s a very common way in to looking at corpus analysis.

Concordance lines, you’ve got an example here on the right. This is taken from a corpus of young adult fiction that I’ve been working on, and here I search for the term twists in the corpus. They’re sorted to the left, so you’ve got words on the left and words on the right. The words on the left are in alphabetic order, so Julie occurs before Leah occurs before Maeve. And these give you insight into how the word twists is used across a variety of texts, not just a single text. So, you read these concordance lines by going downwards. And you can see here that near the bottom you’ve got two occurrences of stomach twists. And when we look at collocates for twist, stomach is quite high on the list here, so that would be another way in to look at collocate lists for words.

To explore a corpus you need corpus software, and here are some currently popular tools. These are all either free or fairly low cost to buy, and if you’re at a university you might have a subscription.

[0:06:00]

Turning now to discourse analysis, the other part of CADS, what do we mean by this? So, discourse is generally defined as meaning making through language above the level of the sentence, so it’s not looking at words and phrases in isolation, but longer stretches of text to see how meaning is built up in context. And discourse analysis is a widely used method across the social sciences.

So, why would you use discourse analysis? Well, generally, it’s used at a broad level to study what humans do with language and how we do it. So, how is meaning built up gradually both within an individual text and across several texts? Do words take on particular meanings through the words surrounding them, the co-text? And this is a qualitative method where you’re closely reading text to gain right insights from a few texts.

So, ways of analysing discourse include the ones listed here. You might read individual texts and form hypotheses, and then read more texts and see if these follow through. You might search for a particular term and extract a set of examples which you then code. This could be within software or qualitative software within Excel. You might thematically analyse language data, so in terms of CADS, once you’ve got your corpus and begun analysing it through corpus linguistics, the discourse analysis side allows you to analyse it further in a closer, more qualitative way. You might also draw on other sources of data alongside the textual data, for example interviews, focus groups, or survey data.

So, taken together, corpus linguistics and discourse analysis form, as I said earlier, quite a powerful combination. From corpus linguistics there are different entry points to examining your whole textual dataset. You could look at frequency lists, collocates or extra keywords, and then rather than simply look at this language in isolation, the CADS analyst would then draw on discourse analytic techniques to qualitatively analyse the data or a sample of the data. So, you’ve got he more objective corpus linguistic techniques as an entry point to avoid any cherry picking of interesting examples, and then the human qualitative analysis to interpret this data.

Alternatively, you might start from qualitative reading of individual texts and notice particular features of the texts and then search for these in a corpus. CADS research tends to be an iterative process, where you may be extract keywords, categorise them qualitatively into thematic groups, search for particular terms in the corpus, and back and forth between the individual texts and the close analysis and the broader corpus techniques.

In short then, combining the two methods provides different ways in or entry points to exploring your dataset, allowing you to look at language in different ways. And this form of triangulation is a powerful way of understanding what people mean by what is said and written, and how meanings are created.

National Centre for Research Methods (NCRM)  
Social Sciences  
Murray Building (Bldg 58)  
University of Southampton  
Southampton SO17 1BJ  
United Kingdom

**Web** www.ncrm.ac.uk   
**Email** info@ncrm.ac.uk  
**Tel** +44 23 8059 4539  
**Twitter** @NCRMUK